Kongeriget Danmark

Patent application No.: PA 2000 00576

Date of filing: 07 April 2000

Applicant: Novozymes A/S
Krogshøjvej 36
DK-2880 Bagsværd

This is to certify the correctness of the following information:

The attached photocopy is a true copy of the following document:

- The specification, claims, abstract, drawings and sequence listing filed with the application on the filing date indicated above.

  By assignment dated 17 Nov 2000 and filed on 01 Dec 2000, the application has been assigned to Novozymes A/S

Patent- og Varemærkestyrelsen Kongeriget Danmark

Patent- og Varemærkestyrelsen
Erhvervsministeriet

TAASTRUP 21 March 2001

Lizzi Vester

Lizzi Vester
Head of Section

Signal sequence trapping

Field of invention

A method for isolating genes encoding secreted polypeptides from existing gene libraries is described in which the endogenous secretion signal sequences are detected using an *in vitro* transposition reaction where the transposon contains a secretion reporter.

Background of the invention

The search for new industrial enzymes and more specifically secreted enzymes is presently reliant on the availability of simple primary functional assays. Typically the substrate is used in the growth medium for the screening of microorganisms and degradation of the substrate may be recognized by a physical change in the substrate (colour change, halo formation around a colony, fluorescence etc.). Many proteins exist for which there is no simple functional assay and these may have potential application as industrial enzymes.

Enzymes which are secreted are highly interesting for use in industrial applications. A positive selection screening system which selects only clones encoding secreted enzymes is thus very desirable. Signal trapping is a method to identify genes containing a signal peptide using a translational fusion to an extracellular reporter gene lacking its own signal. This has been reported in the literature for the purpose of identifying new signal sequences (Smith, H. et al., 1987, Construction and use of signal sequence selection vectors in *Escherichia coli* and *Bacillus subtilis*. J. Bact. 169:3321-3328), also the use of such for defining clearly the specific elements within signal peptides which are required for optimal function (Smith, H. et al, 1988. Characterisation of signal-sequence-coding regions selected from the *Bacillus subtilis* chromosome. Gene. 70:351-361).

A number of publications describe cloning vector reporter systems where genomic or cDNA libraries are constructed in a screening vector containing a signal-less reporter gene. When a cDNA or genomic fragment lacking a translational stop site is cloned upstream of the reporter gene in a translational fusion, a resulting protein-reporter gene fusion product is formed. If the cDNA or genomic fragment cloned contains a signal peptide, the fusion protein is secreted to the outside of the cell. Secretion can be detected by growth on selective media as in the use of invertase in *Saccharomyces cerevisiae* or in the use of eg. β-lactamase in *Escherichia coli*. These publications are not concerned with methods for screening previously established gene libraries.

The number of clones to be investigated in the library is dramatically reduced by the screening to those containing a signal peptide, however a resulting clone may only contain an

incomplete gene which may or may not include the minimum DNA information needed to encode the enzymatic activity originally associated with the secretion signal sequence isolated.

Summary of the Invention

The problem to be solved by the present invention is to identify those clones in an existing gene library that encode efficiently secreted polypeptides or enzymes, even enzymes with unknown activity, without having to reclone a library in a screening-vector and without having to screen the library in traditional labour- and time consuming enzyme activity assays that would detect known enzyme activities only. Solving this problem allows rapid and efficient industrial exploitation of relevant secreted polypeptides from new organisms from which gene libraries were previously established for another purpose, even enzymes for which the specific substrate is unknown can be found.

We describe the combination of the use of a signal-less reporter gene and a random *in vitro* transposition reaction for the identification of genes encoding secreted polypeptides from genomic or cDNA libraries previously established, *e.g.* the use of a signal-less β-lactamase gene in a transposon such as the MuA transposon. The present invention allows to screen previously established genebanks or libraries by proxy for genes encoding secreted polypeptides or enzymes which would likely not have been isolated using conventional screening assays.

Accordingly in a first aspect the invention relates to a method for identifying and isolating a gene of interest from a gene library, wherein said gene encodes a polypeptide carrying a secretion signal sequence, the method comprising the steps of:

a) providing a genomic DNA library or a cDNA library;
b) inserting randomly into said library a DNA fragment comprising a promoterless and secretion signal-less gene encoding a secretion reporter;
c) introducing the library carrying random inserts of said DNA fragment into a population of host cells;
d) screening for a host cell that expresses and secretes the secretion reporter;
e) identifying the gene of interest into which the secretion reporter was inserted by sequencing the DNA flanking the DNA fragment of step b; and
f) isolating the complete gene of interest from the library of step a).

In a non-limiting example herein existing cDNA or genomic DNA libraries are tagged with a transposon containing a reporter gene, all in-frame fusions with a gene containing a signal sequence will result in clone growth. The upstream and downstream areas of transposon insertion are then sequenced and the gene is identified by sequence analysis. In many cases, obtaining the full sequence of a tagged gene will be facilitated by the recovery of

numerous clones of the same gene tagged at different sites. Positive clones are sequenced to identify clones that represent the same gene but have different transposon insertion sites. In this way all or most of the open reading frame (ORF) can be obtained by contig assembly. If a complete ORF is not obtained, then the full length gene may be obtained by primer walking.

The sequence information thus obtained can then be used to isolate the complete gene including the sequence encoding the secretion signal sequence and further to make an optimal expression construct for industrial production of the secreted proteins, all well within the skill of the art, whereafter the industrial production process of the enzyme is a matter thoroughly described in the art as shown elsewhere herein.

Accordingly in a second aspect the invention relates to a gene of interest isolated from a gene library, wherein said gene encodes a protein carrying a secretion signal sequence, and wherein said gene is isolated by the method of the present invention.

In a third aspect the invention relates to an enzyme encoded by a gene of interest as defined in the previous aspect.

Further in another aspect the invention relates to an expression system comprising a gene of interest as defined in the second aspect.

Yet other aspects of the invention relate to a host cell comprising an expression system as defined in the previous aspect, or to a host cell comprising at least two chromosomally integrated copies of a gene of interest as defined in the third aspect.

In a final aspect the invention relates to a process for producing an enzyme comprising cultivating a host cell as defined in the previous aspects under conditions suitable for expressing a gene of interest as defined in the third aspect, wherein said host cell secretes a protein encoded by said gene into the growht medium.

Drawings

Figure 1. Plasmid map of a construct containing the transposon MuA and the signal-less β-lactamase gene of the non-limiting example 1 of the present application.

Figure 2. Schematic for prokaryotic transposon signal trapping system.

Definitions

In accordance with the present invention there may be employed conventional molecular biology, microbiology, and recombinant DNA techniques within the skill of the art. Such techniques are explained fully in the literature. See, e.g., Sambrook, Fritsch & Maniatis, *Molecular Cloning: A Laboratory Manual*, Second Edition (1989) Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York (herein "Sambrook et al., 1989") *DNA Cloning: A Practical Approach*, Volumes I and II /D.N. Glover ed. 1985); *Oligonucleotide Synthesis* (M.J. Gait ed. 1984); *Nucleic Acid Hybridization* (B.D. Hames & S.J. Higgins eds

ed. 1984); *Nucleic Acid Hybridization* (B.D. Hames & S.J. Higgins eds (1985)); *Transcription And Translation* (B.D. Hames & S.J. Higgins, eds. (1984)); *Animal Cell Culture* (R.I. Freshney, ed. (1986)); *Immobilized Cells And Enzymes* (IRL Press, (1986)); B. Perbal, *A Practical Guide To Molecular Cloning* (1984).

When applied to a protein, the term "isolated" indicates that the protein is found in a condition other than its native environment, such as apart from blood and animal tissue. In a preferred form, the isolated protein is substantially free of other proteins, particularly other proteins of animal origin. It is preferred to provide the proteins in a highly purified form, i.e., greater than 95% pure, more preferably greater than 99% pure. When applied to a polynucleotide molecule, the term "isolated" indicates that the molecule is removed from its natural genetic milieu, and is thus free of other extraneous or unwanted coding sequences, and is in a form suitable for use within genetically engineered protein production systems. Such isolated molecules are those that are separated from their natural environment and include cDNA and genomic clones. Isolated DNA molecules of the present invention are free of other genes with which they are ordinarily associated, and may include naturally occurring 5' and 3' untranslated regions such as promoters and terminators. The identification of associated regions will be evident to one of ordinary skill in the art (see for example, Dynan and Tijan, Nature 316: 774-78, 1985).

A "polynucleotide" is a single- or double-stranded polymer of deoxyribonucleotide or ribonucleotide bases read from the 5' to the 3' end. Polynucleotides include RNA and DNA, and may be isolated from natural sources, synthesized *in vitro*, or prepared from a combination of natural and synthetic molecules. A "nucleic acid molecule" refers to the phosphate ester polymeric form of ribonucleosides (adenosine, guanosine, uridine or cytidine; "RNA molecules") or deoxyribonucleosides (deoxyadenosine, deoxyguanosine, deoxythymidine, or deoxycytidine; "DNA molecules") in either single stranded form, or a double-stranded helix. Double stranded DNA-DNA, DNA-RNA and RNA-RNA helices are possible. The term nucleic acid molecule, and in particular DNA or RNA molecule, refers only to the primary and secondary structure of the molecule, and does not limit it to any particular tertiary or quaternary forms. Thus, this term includes double-stranded DNA found, *inter alia*, in linear or circular DNA molecules (e.g., restriction fragments), plasmids, and chromosomes. In discussing the structure of particular double-stranded DNA molecules, sequences may be described herein according to the normal convention of giving only the sequence in the 5' to 3' direction along the nontranscribed strand of DNA (i.e., the strand having a sequence homologous to the mRNA). A "recombinant DNA molecule" is a DNA molecule that has undergone a molecular biological manipulation.

Nucleic Acid Constructs

The present invention also relates to nucleic acid constructs comprising a nucleic acid sequence of the present invention operably linked to one or more control sequences which direct the expression of the coding sequence in a suitable host cell under conditions compatible with the control sequences. Expression will be understood to include any step involved in the production of the polypeptide including, but not limited to, transcription, post-transcriptional modification, translation, post-translational modification, and secretion.

"Expression construct" is defined herein as a nucleic acid molecule, either single- or double-stranded, which is isolated from a naturally occurring gene or which has been modified to contain segments of nucleic acid combined and juxtaposed in a manner that would not otherwise exist in nature. The term nucleic acid construct is synonymous with the term expression cassette when the nucleic acid construct contains all the control sequences required for expression of a coding sequence of the present invention. The term "coding sequence" is defined herein as a nucleic acid sequence which directly specifies the amino acid sequence of its protein product. The boundaries of the coding sequence are generally determined by a ribosome binding site (prokaryotes) or by the ATG start codon (eukaryotes) located just upstream of the open reading frame at the 5' end of the mRNA and a transcription terminator sequence located just downstream of the open reading frame at the 3' end of the mRNA. A coding sequence can include, but is not limited to, DNA, cDNA, and recombinant nucleic acid sequences.

An isolated nucleic acid sequence encoding a polypeptide of the present invention may be manipulated in a variety of ways to provide for expression of the polypeptide. Manipulation of the nucleic acid sequence prior to its insertion into a vector may be desirable or necessary depending on the expression vector. The techniques for modifying nucleic acid sequences utilizing recombinant DNA methods are well known in the art.

The term "control sequences" is defined herein to include all components which are necessary or advantageous for the expression of a polypeptide of the present invention. Each control sequence may be native or foreign to the nucleic acid sequence encoding the polypeptide. Such control sequences include, but are not limited to, a leader, polyadenylation sequence, propeptide sequence, promoter, signal peptide sequence, and transcription terminator. At a minimum, the control sequences include a promoter, and transcriptional and translational stop signals. The control sequences may be provided with linkers for the purpose of introducing specific restriction sites facilitating ligation of the control sequences with the coding region of the nucleic acid sequence encoding a polypeptide. The term "operably linked" is defined herein as a configuration in which a control sequence is appropriately placed at a position relative to the coding sequence of the DNA sequence such that the control sequence directs the expression of a polypeptide.

The control sequence may be an appropriate promoter sequence, a nucleic acid sequence which is recognized by a host cell for expression of the nucleic acid sequence. The promoter sequence contains transcriptional control sequences which mediate the expression of the polypeptide. The promoter may be any nucleic acid sequence which shows transcriptional activity in the host cell of choice including mutant, truncated, and hybrid promoters, and may be obtained from genes encoding extracellular or intracellular polypeptides either homologous or heterologous to the host cell.

Examples of suitable promoters for directing the transcription of the nucleic acid constructs of the present invention, especially in a bacterial host cell, are the promoters obtained from the *E. coli lac* operon, *Streptomyces coelicolor* agarase gene (*dagA*), *Bacillus subtilis* levansucrase gene (*sacB*), *Bacillus licheniformis* alpha-amylase gene (*amyL*), *Bacillus stearothermophilus* maltogenic amylase gene (*amyM*), *Bacillus amyloliquefaciens* alpha-amylase gene (*amyQ*), *Bacillus licheniformis* penicillinase gene (*penP*), *Bacillus subtilis xylA* and *xylB* genes, and prokaryotic beta-lactamase gene (Villa-Kamaroff *et al.*, 1978, *Proceedings of the National Academy of Sciences USA* 75: 3727-3731), as well as the *tac* promoter (DeBoer *et al.*, 1983, *Proceedings of the National Academy of Sciences USA* 80: 21-25). Further promoters are described in "Useful proteins from recombinant bacteria" in *Scientific American*, 1980, 242: 74-94; and in Sambrook, J. et al., 1989, *Molecular Cloning, A Laboratory Manual*, 2d edition, Cold Spring Harbor, New York.

Examples of suitable promoters for directing the transcription of the nucleic acid constructs of the present invention in a filamentous fungal host cell are promoters obtained from the genes for *Aspergillus oryzae* TAKA amylase, *Rhizomucor miehei* aspartic proteinase, *Aspergillus niger* neutral alpha-amylase, *Aspergillus niger* acid stable alpha-amylase, *Aspergillus niger* or *Aspergillus awamori* glucoamylase (*glaA*), *Rhizomucor miehei* lipase, *Aspergillus oryzae* alkaline protease, *Aspergillus oryzae* triose phosphate isomerase, *Aspergillus nidulans* acetamidase, and *Fusarium oxysporum* trypsin-like protease (WO 96/00787), as well as the NA2-tpi promoter (a hybrid of the promoters from the genes for *Aspergillus niger* neutral alpha-amylase and *Aspergillus oryzae* triose phosphate isomerase), and mutant, truncated, and hybrid promoters thereof.

In a yeast host, useful promoters are obtained from the genes for *Saccharomyces cerevisiae* enolase (ENO-1), *Saccharomyces cerevisiae* galactokinase (GAL1), *Saccharomyces cerevisiae* alcohol dehydrogenase/glyceraldehyde-3-phosphate dehydrogenase (ADH2/GAP), and *Saccharomyces cerevisiae* 3-phosphoglycerate kinase. Other useful promoters for yeast host cells are described by Romanos *et al.*, 1992, *Yeast* 8: 423-488.

The control sequence may also be a suitable transcription terminator sequence, a sequence recognized by a host cell to terminate transcription. The terminator sequence is oper-

ably linked to the 3' terminus of the nucleic acid sequence encoding the polypeptide. Any terminator which is functional in the host cell of choice may be used in the present invention.

Preferred terminators for filamentous fungal host cells are obtained from the genes for *Aspergillus oryzae* TAKA amylase, *Aspergillus niger* glucoamylase, *Aspergillus nidulans* anthranilate synthase, *Aspergillus niger* alpha-glucosidase, and *Fusarium oxysporum* trypsin-like protease.

Preferred terminators for yeast host cells are obtained from the genes for *Saccharomyces cerevisiae* enolase, *Saccharomyces cerevisiae* cytochrome C (CYC1), and *Saccharomyces cerevisiae* glyceraldehyde-3-phosphate dehydrogenase. Other useful terminators for yeast host cells are described by Romanos *et al.*, 1992, *supra*.

The control sequence may also be a suitable leader sequence, a nontranslated region of an mRNA which is important for translation by the host cell. The leader sequence is operably linked to the 5' terminus of the nucleic acid sequence encoding the polypeptide. Any leader sequence that is functional in the host cell of choice may be used in the present invention.

Preferred leaders for filamentous fungal host cells are obtained from the genes for *Aspergillus oryzae* TAKA amylase and *Aspergillus nidulans* triose phosphate isomerase.

Suitable leaders for yeast host cells are obtained from the genes for *Saccharomyces cerevisiae* enolase (ENO-1), *Saccharomyces cerevisiae* 3-phosphoglycerate kinase, *Saccharomyces cerevisiae* alpha-factor, and *Saccharomyces cerevisiae* alcohol dehydrogenase/glyceraldehyde-3-phosphate dehydrogenase (ADH2/GAP).

The control sequence may also be a polyadenylation sequence, a sequence operably linked to the 3' terminus of the nucleic acid sequence and which, when transcribed, is recognized by the host cell as a signal to add polyadenosine residues to transcribed mRNA. Any polyadenylation sequence which is functional in the host cell of choice may be used in the present invention.

Preferred polyadenylation sequences for filamentous fungal host cells are obtained from the genes for *Aspergillus oryzae* TAKA amylase, *Aspergillus niger* glucoamylase, *Aspergillus nidulans* anthranilate synthase, *Fusarium oxysporum* trypsin-like protease, and *Aspergillus niger* alpha-glucosidase.

Useful polyadenylation sequences for yeast host cells are described by Guo and Sherman, 1995, *Molecular Cellular Biology* 15: 5983-5990.

It may also be desirable to add regulatory sequences which allow the regulation of the expression of the polypeptide relative to the growth of the host cell. Examples of regulatory systems are those which cause the expression of the gene to be turned on or off in response to a chemical or physical stimulus, including the presence of a regulatory compound. Regulatory systems in prokaryotic systems include the *lac*, *tac*, and *trp* operator systems. In yeast,

the ADH2 system or GAL1 system may be used. In filamentous fungi, the TAKA alpha-amylase promoter, *Aspergillus niger* glucoamylase promoter, and *Aspergillus oryzae* glucoamylase promoter may be used as regulatory sequences. Other examples of regulatory sequences are those which allow for gene amplification. In eukaryotic systems, these include the dihydrofolate reductase gene which is amplified in the presence of methotrexate, and the metallothionein genes which are amplified with heavy metals. In these cases, the nucleic acid sequence encoding the polypeptide would be operably linked with the regulatory sequence.

The present invention also relates to nucleic acid constructs for altering the expression of an endogenous gene encoding a polypeptide of the present invention. The constructs may contain the minimal number of components necessary for altering expression of the endogenous gene. In one embodiment, the nucleic acid constructs preferably contain (a) a targeting sequence, (b) a regulatory sequence, (c) an exon, and (d) a splice-donor site. Upon introduction of the nucleic acid construct into a cell, the construct inserts by homologous recombination into the cellular genome at the endogenous gene site. The targeting sequence directs the integration of elements (a)-(d) into the endogenous gene such that elements (b)-(d) are operably linked to the endogenous gene. In another embodiment, the nucleic acid constructs contain (a) a targeting sequence, (b) a regulatory sequence, (c) an exon, (d) a splice-donor site, (e) an intron, and (f) a splice-acceptor site, wherein the targeting sequence directs the integration of elements (a)-(f) such that elements (b)-(f) are operably linked to the endogenous gene. However, the constructs may contain additional components such as a selectable marker.

The introduction of these components results in production of a new transcription unit in which expression of the endogenous gene is altered. In essence, the new transcription unit is a fusion product of the sequences introduced by the targeting constructs and the endogenous gene. In one embodiment in which the endogenous gene is altered, the gene is activated. In this embodiment, homologous recombination is used to replace, disrupt, or disable the regulatory region normally associated with the endogenous gene of a parent cell through the insertion of a regulatory sequence which causes the gene to be expressed at higher levels than evident in the corresponding parent cell.

The constructs further contain one or more exons of the endogenous gene. An exon is defined as a DNA sequence which is copied into RNA and is present in a mature mRNA molecule such that the exon sequence is in-frame with the coding region of the endogenous gene. The exons can, optionally, contain DNA which encodes one or more amino acids and/or partially encodes an amino acid. Alternatively, the exon contains DNA which corresponds to a 5' non-encoding region. Where the exogenous exon or exons encode one or more amino acids and/or a portion of an amino acid, the nucleic acid construct is designed such that, upon transcription and splicing, the reading frame is in-frame with the coding region of the endoge-

nous gene so that the appropriate reading frame of the portion of the mRNA derived from the second exon is unchanged.

The splice-donor site of the constructs directs the splicing of one exon to another exon. Typically, the first exon lies 5' of the second exon, and the splice-donor site overlapping and flanking the first exon on its 3' side recognizes a splice-acceptor site flanking the second exon on the 5' side of the second exon. A splice-acceptor site, like a splice-donor site, is a sequence which directs the splicing of one exon to another exon. Acting in conjunction with a splice-donor site, the splicing apparatus uses a splice-acceptor site to effect the removal of an intron.

Expression Vectors

The present invention also relates to recombinant expression vectors comprising a nucleic acid sequence of the present invention, a promoter, and transcriptional and translational stop signals. The various nucleic acid and control sequences described above may be joined together to produce a recombinant expression vector which may include one or more convenient restriction sites to allow for insertion or substitution of the nucleic acid sequence encoding the polypeptide at such sites. Alternatively, the nucleic acid sequence of the present invention may be expressed by inserting the nucleic acid sequence or a nucleic acid construct comprising the sequence into an appropriate vector for expression. In creating the expression vector, the coding sequence is located in the vector so that the coding sequence is operably linked with the appropriate control sequences for expression.

The recombinant expression vector may be any vector (*e.g.*, a plasmid or virus) which can be conveniently subjected to recombinant DNA procedures and can bring about the expression of the nucleic acid sequence. The choice of the vector will typically depend on the compatibility of the vector with the host cell into which the vector is to be introduced. The vectors may be linear or closed circular plasmids.

The vector may be an autonomously replicating vector, *i.e.*, a vector which exists as an extrachromosomal entity, the replication of which is independent of chromosomal replication, *e.g.*, a plasmid, an extrachromosomal element, a minichromosome, or an artificial chromosome. The vector may contain any means for assuring self-replication. Alternatively, the vector may be one which, when introduced into the host cell, is integrated into the genome and replicated together with the chromosome(s) into which it has been integrated. Furthermore, a single vector or plasmid or two or more vectors or plasmids which together contain the total DNA to be introduced into the genome of the host cell, or a transposon may be used.

The vectors of the present invention preferably contain one or more selectable markers which permit easy selection of transformed cells. A selectable marker is a gene the product of which provides for biocide or viral resistance, resistance to heavy metals, prototrophy to

auxotrophs, and the like. Examples of bacterial selectable markers are the *dal* genes from *Bacillus subtilis* or *Bacillus licheniformis*, or markers which confer antibiotic resistance such as ampicillin, kanamycin, chloramphenicol or tetracycline resistance. Suitable markers for yeast host cells are ADE2, HIS3, LEU2, LYS2, MET3, TRP1, and URA3. Selectable markers for use in a filamentous fungal host cell include, but are not limited to, *amdS* (acetamidase), *argB* (ornithine carbamoyltransferase), *bar* (phosphinothricin acetyltransferase), *hygB* (hygromycin phosphotransferase), *niaD* (nitrate reductase), *pyrG* (orotidine-5'-phosphate decarboxylase), *sC* (sulfate adenyltransferase), *trpC* (anthranilate synthase), as well as equivalents thereof. Preferred for use in an *Aspergillus* cell are the *amdS* and *pyrG* genes of *Aspergillus nidulans* or *Aspergillus oryzae* and the *bar* gene of *Streptomyces hygroscopicus*.

The vectors of the present invention preferably contain an element(s) that permits stable integration of the vector into the host cell's genome or autonomous replication of the vector in the cell independent of the genome.

For integration into the host cell genome, the vector may rely on the nucleic acid sequence encoding the polypeptide or any other element of the vector for stable integration of the vector into the genome by homologous or nonhomologous recombination. Alternatively, the vector may contain additional nucleic acid sequences for directing integration by homologous recombination into the genome of the host cell. The additional nucleic acid sequences enable the vector to be integrated into the host cell genome at a precise location(s) in the chromosome(s). To increase the likelihood of integration at a precise location, the integrational elements should preferably contain a sufficient number of nucleic acids, such as 100 to 1,500 base pairs, preferably 400 to 1,500 base pairs, and most preferably 800 to 1,500 base pairs, which are highly homologous with the corresponding target sequence to enhance the probability of homologous recombination. The integrational elements may be any sequence that is homologous with the target sequence in the genome of the host cell. Furthermore, the integrational elements may be non-encoding or encoding nucleic acid sequences. On the other hand, the vector may be integrated into the genome of the host cell by non-homologous recombination.

For autonomous replication, the vector may further comprise an origin of replication enabling the vector to replicate autonomously in the host cell in question. Examples of bacterial origins of replication are the origins of replication of plasmids pBR322, pUC19, pACYC177, and pACYC184 permitting replication in *E. coli*, and pUB110, pE194, pTA1060, and pAMß1 permitting replication in *Bacillus*. Examples of origins of replication for use in a yeast host cell are the 2 micron origin of replication, ARS1, ARS4, the combination of ARS1 and CEN3, and the combination of ARS4 and CEN6. The origin of replication may be one having a mutation which makes its functioning temperature-sensitive in the host cell (see, *e.g.*, Ehrlich, 1978, *Proceedings of the National Academy of Sciences USA* 75: 1433).

More than one copy of a nucleic acid sequence of the present invention may be inserted into the host cell to increase production of the gene product. An increase in the copy number of the nucleic acid sequence can be obtained by integrating at least one additional copy of the sequence into the host cell genome or by including an amplifiable selectable marker gene with the nucleic acid sequence where cells containing amplified copies of the selectable marker gene, and thereby additional copies of the nucleic acid sequence, can be selected for by cultivating the cells in the presence of the appropriate selectable agent.

The procedures used to ligate the elements described above to construct the recombinant expression vectors of the present invention are well known to one skilled in the art (see, *e.g.*, Sambrook *et al.*, 1989, *supra*).

Host Cells

The present invention also relates to recombinant host cells, comprising a nucleic acid sequence of the invention, which are advantageously used in the recombinant production of the polypeptides. A vector comprising a nucleic acid sequence of the present invention is introduced into a host cell so that the vector is maintained as a chromosomal integrant or as a self-replicating extra-chromosomal vector as described earlier. The term "host cell" encompasses any progeny of a parent cell that is not identical to the parent cell due to mutations that occur during replication. The choice of a host cell will to a large extent depend upon the gene encoding the polypeptide and its source.

The host cell may be a unicellular microorganism, *e.g.*, a prokaryote, or a non-unicellular microorganism, *e.g.*, a eukaryote.

Useful unicellular cells are bacterial cells such as gram positive bacteria including, but not limited to, a *Bacillus* cell, *e.g.*, *Bacillus alkalophilus*, *Bacillus amyloliquefaciens*, *Bacillus brevis*, *Bacillus circulans*, *Bacillus clausii*, *Bacillus coagulans*, *Bacillus lautus*, *Bacillus lentus*, *Bacillus licheniformis*, *Bacillus megaterium*, *Bacillus stearothermophilus*, *Bacillus subtilis*, and *Bacillus thuringiensis*; or a *Streptomyces* cell, *e.g.*, *Streptomyces lividans* or *Streptomyces murinus*, or gram negative bacteria such as *E. coli* and *Pseudomonas* sp. In a preferred embodiment, the bacterial host cell is a *Bacillus lentus*, *Bacillus licheniformis*, *Bacillus stearothermophilus*, or *Bacillus subtilis* cell. In another preferred embodiment, the *Bacillus* cell is an alkalophilic *Bacillus*.

The introduction of a vector into a bacterial host cell may, for instance, be effected by protoplast transformation (see, *e.g.*, Chang and Cohen, 1979, *Molecular General Genetics* 168: 111-115), using competent cells (see, *e.g.*, Young and Spizizin, 1961, *Journal of Bacteriology* 81: 823-829, or Dubnau and Davidoff-Abelson, 1971, *Journal of Molecular Biology* 56: 209-221), electroporation (see, *e.g.*, Shigekawa and Dower, 1988, *Biotechniques* 6: 742-751), or conjugation (see, *e.g.*, Koehler and Thorne, 1987, *Journal of Bacteriology* 169: 5771-5278).

The host cell may be a eukaryote, such as a mammalian, insect, plant, or fungal cell.

In a preferred embodiment, the host cell is a fungal cell. "Fungi" as used herein includes the phyla Ascomycota, Basidiomycota, Chytridiomycota, and Zygomycota (as defined by Hawksworth *et al.*, *In, Ainsworth and Bisby's Dictionary of The Fungi*, 8th edition, 1995, CAB International, University Press, Cambridge, UK) as well as the Oomycota (as cited in Hawksworth *et al.*, 1995, *supra*, page 171) and all mitosporic fungi (Hawksworth *et al.*, 1995, *supra*).

In a more preferred embodiment, the fungal host cell is a yeast cell. "Yeast" as used herein includes ascosporogenous yeast (Endomycetales), basidiosporogenous yeast, and yeast belonging to the Fungi Imperfecti (Blastomycetes). Since the classification of yeast may change in the future, for the purposes of this invention, yeast shall be defined as described in *Biology and Activities of Yeast* (Skinner, F.A., Passmore, S.M., and Davenport, R.R., eds, *Soc. App. Bacteriol. Symposium Series* No. 9, 1980).

In an even more preferred embodiment, the yeast host cell is a *Candida*, *Hansenula*, *Kluyveromyces*, *Pichia*, *Saccharomyces*, *Schizosaccharomyces*, or *Yarrowia* cell.

In a most preferred embodiment, the yeast host cell is a *Saccharomyces carlsbergensis*, *Saccharomyces cerevisiae*, *Saccharomyces diastaticus*, *Saccharomyces douglasii*, *Saccharomyces kluyveri*, *Saccharomyces norbensis* or *Saccharomyces oviformis* cell. In another most preferred embodiment, the yeast host cell is a *Kluyveromyces lactis* cell. In another most preferred embodiment, the yeast host cell is a *Yarrowia lipolytica* cell.

In another more preferred embodiment, the fungal host cell is a filamentous fungal cell. "Filamentous fungi" include all filamentous forms of the subdivision Eumycota and Oomycota (as defined by Hawksworth *et al.*, 1995, *supra*). The filamentous fungi are characterized by a mycelial wall composed of chitin, cellulose, glucan, chitosan, mannan, and other complex polysaccharides. Vegetative growth is by hyphal elongation and carbon catabolism is obligately aerobic. In contrast, vegetative growth by yeasts such as *Saccharomyces cerevisiae* is by budding of a unicellular thallus and carbon catabolism may be fermentative.

In an even more preferred embodiment, the filamentous fungal host cell is a cell of a species of, but not limited to, *Acremonium*, *Aspergillus*, *Fusarium*, *Humicola*, *Mucor*, *Myceliophthora*, *Neurospora*, *Penicillium*, *Thielavia*, *Tolypocladium*, or *Trichoderma*.

In a most preferred embodiment, the filamentous fungal host cell is an *Aspergillus awamori*, *Aspergillus foetidus*, *Aspergillus japonicus*, *Aspergillus nidulans*, *Aspergillus niger* or *Aspergillus oryzae* cell. In another most preferred embodiment, the filamentous fungal host cell is a *Fusarium bactridioides*, *Fusarium cerealis*, *Fusarium crookwellense*, *Fusarium culmorum*, *Fusarium graminearum*, *Fusarium graminum*, *Fusarium heterosporum*, *Fusarium negundi*, *Fusarium oxysporum*, *Fusarium reticulatum*, *Fusarium roseum*, *Fusarium sambucinum*, *Fusarium sarcochroum*, *Fusarium sporotrichioides*, *Fusarium sulphureum*, *Fusarium torulo-*

sum, *Fusarium trichothecioides*, or *Fusarium venenatum* cell. In an even most preferred embodiment, the filamentous fungal parent cell is a *Fusarium venenatum* (Nirenberg sp. nov.) cell. In another most preferred embodiment, the filamentous fungal host cell is a *Humicola insolens*, *Humicola lanuginosa*, *Mucor miehei*, *Myceliophthora thermophila*, *Neurospora crassa*, *Penicillium purpurogenum*, *Thielavia terrestris*, *Trichoderma harzianum*, *Trichoderma koningii*, *Trichoderma longibrachiatum*, *Trichoderma reesei*, or *Trichoderma viride* cell.

Fungal cells may be transformed by a process involving protoplast formation, transformation of the protoplasts, and regeneration of the cell wall in a manner known *per se*. Suitable procedures for transformation of *Aspergillus* host cells are described in EP 238 023 and Yelton *et al.*, 1984, *Proceedings of the National Academy of Sciences USA* 81: 1470-1474. Suitable methods for transforming *Fusarium* species are described by Malardier *et al.*, 1989, *Gene* 78: 147-156 and WO 96/00787. Yeast may be transformed using the procedures described by Becker and Guarente, *In* Abelson, J.N. and Simon, M.I., editors, *Guide to Yeast Genetics and Molecular Biology*, *Methods in Enzymology*, Volume 194, pp 182-187, Academic Press, Inc., New York; Ito *et al.*, 1983, *Journal of Bacteriology* 153: 163; and Hinnen *et al.*, 1978, *Proceedings of the National Academy of Sciences USA* 75: 1920.

Process of Production

The present invention also relates to processes for producing a polypeptide of the present invention comprising (a) cultivating a strain, which in its wild-type form is capable of producing the polypeptide, to produce a supernatant comprising the polypeptide; and (b) recovering the polypeptide.

The present invention further relates to methods for producing a polypeptide of the present invention comprising (a) cultivating a homologously recombinant cell, having incorporated therein a new transcription unit comprising a regulatory sequence, an exon, and/or a splice donor site operably linked to a second exon of an endogenous nucleic acid sequence encoding the polypeptide, under conditions conducive for production of the polypeptide; and (b) recovering the polypeptide. The methods are based on the use of gene activation technology, for example, as described in U.S. Patent No. 5,641,670.

In the production methods of the present invention, the cells are cultivated in a nutrient medium suitable for production of the polypeptide using methods known in the art. For example, the cell may be cultivated by shake flask cultivation, small-scale or large-scale fermentation (including continuous, batch, fed-batch, or solid state fermentations) in laboratory or industrial fermentors performed in a suitable medium and under conditions allowing the polypeptide to be expressed and/or isolated. The cultivation takes place in a suitable nutrient medium comprising carbon and nitrogen sources and inorganic salts, using procedures known in the art. Suitable media are available from commercial suppliers or may be prepared according to

published compositions (*e.g.*, in catalogues of the American Type Culture Collection). If the polypeptide is secreted into the nutrient medium, the polypeptide can be recovered directly from the medium. If the polypeptide is not secreted, it can be recovered from cell lysates.

The polypeptides may be detected using methods known in the art that are specific for the polypeptides. These detection methods may include use of specific antibodies, formation of an enzyme product, or disappearance of an enzyme substrate. For example, an enzyme assay may be used to determine the activity of the polypeptide as described herein.

The resulting polypeptide may be recovered by methods known in the art. For example, the polypeptide may be recovered from the nutrient medium by conventional procedures including, but not limited to, centrifugation, filtration, extraction, spray-drying, evaporation, or precipitation.

The polypeptides of the present invention may be purified by a variety of procedures known in the art including, but not limited to, chromatography (*e.g.*, ion exchange, affinity, hydrophobic, chromatofocusing, and size exclusion), electrophoretic procedures (*e.g.*, preparative isoelectric focusing), differential solubility (*e.g.*, ammonium sulfate precipitation), SDS-PAGE, or extraction (see, *e.g.*, *Protein Purification*, J.-C. Janson and Lars Ryden, editors, VCH Publishers, New York, 1989).

Detailed description of the invention

The present invention allows to screen previously established genebanks or libraries by proxy for genes encoding secreted polypeptides or enzymes even of unknown activity and thus without known screening assays, polypeptides that may turn out to be of industrial interest and which would likely not have been isolated using conventional screening assays.

A method for identifying and isolating a gene of interest from a gene library, wherein said gene encodes a polypeptide carrying a secretion signal sequence, the method comprising the steps of:

a) providing a genomic DNA library or a cDNA library;

b) inserting randomly into said library a DNA fragment comprising a promoterless and secretion signal-less gene encoding a secretion reporter;

c) introducing the library carrying random inserts of said DNA fragment into a population of host cells;

d) screening for a host cell that expresses and secretes the secretion reporter;

e) identifying the gene of interest into which the secretion reporter was inserted by sequencing the DNA flanking the DNA fragment of step b; and

f) isolating the complete gene of interest from the library of step a).

In the art several ways of inserting a DNA fragment randomly into a genome are known such as transposition, however this usually requires time- and labour consuming mating experiments to be carried out. The present invention can be performed with ease using *in vitro* protocols commercially available as exemplified herein.

Accordingly a preferred embodiment of the present invention relates to a method of the first aspect, wherein step b) is performed *in vitro*.

The present invention can be performed using any gene libraries known in the art, specifically it can also be used with gene libraries of viable but non-culturable organisms as typically seen in environmental samples. Processes of producing representative or normalized gene-libraries from environmental samples containing non culturable organisms have been described in the art (US 5,763,239).

A preferred embodiment of the invention relates to the method of the first aspect, wherein the cDNA or the cDNA library is normalized.

Another preferred embodiment of the invention relates to the method of the first aspect, wherein the genomic DNA library or cDNA library is derived from a microorganism, preferably the microorganism is a fungus, a filamentous fungus or a yeast, even more preferably the microorganism is a bacterium, or most preferably the microorganism is an archaeon.

As described elsewhere herein several methods exist in the art for random integration of DNA framents into larger DNA sequences, one preferred embodiment of the invention relates to the method of the first aspect, wherein the DNA fragment of the first aspect is a transposon, preferably a MuA transposon.

A preferred embodiment of the invention relates to the method of the first aspect, wherein the secretion reporter is a protein which, when secreted from the host cells, allows said cells to grow in the presence of a substance which otherwise inhibits growth of said cells, preferably the secretion reporter is a β-lactamase or an invertase.

For the present invention several host cells can be imagined to work well, the only criterion being that the host cell recognizes the secretion signal sequence of the gene of interest, and that the host cell is capable of synthesizing a functional secretion reporter.

A preferred embodiment of the present invention relates to the method of the first aspect, wherein the host cells are bacterial, preferably the bacterial cells are *Bacillus* cells, more preferably *Bacillus alkalophilus*, *Bacillus amyloliquefaciens*, *Bacillus brevis*, *Bacillus circulans*, *Bacillus clausii*, *Bacillus coagulans*, *Bacillus lautus*, *Bacillus lentus*, *Bacillus licheniformis*, *Bacillus megaterium*, *Bacillus stearothermophilus*, *Bacillus subtilis*, or *Bacillus thuringiensis*.

A preferred embodiment of the present invention relates to the method of the first aspect, wherein the host cells are fungal, preferably the fungal cells are *Candida*, *Kluyveromyces*, *Pichia*, *Saccharomyces*, *Schizosaccharomyces*, *Yarrowia*, *Acremonium*, *Aspergillus*, *Aureobasidium*, *Cryptococcus*, *Filibasidium*, *Fusarium*, *Humicola*, *Magnaporthe*, *Mucor*, *Myceliophthora*, *Neocallimastix*, *Neurospora*, *Paecilomyces*, *Penicillium*, *Piromyces*, *Schizophyllum*, *Talaromyces*, *Thermoascus*, *Thielavia*, *Tolypocladium*, or *Trichoderma*, more preferably the fungal host cells are *Saccharomyces cerevisiae*, *Aspergillus aculeatus*, *Aspergillus awamori*, *Aspergillus nidulans*, *Aspergillus niger*, or *Aspergillus oryzae*.

The fungal host cells of the invention may be *Saccharomyces carlsbergensis*, *Saccharomyces cerevisiae*, *Saccharomyces diastaticus*, *Saccharomyces douglasii*, *Saccharomyces kluyveri*, *Saccharomyces norbensis*, *Saccharomyces oviformis*, *Aspergillus aculeatus*, *Aspergillus awamori*, *Aspergillus foetidus*, *Aspergillus japonicus*, *Aspergillus nidulans*, *Aspergillus niger*, *Aspergillus oryzae*, *Fusarium bactridioides*, *Fusarium cerealis*, *Fusarium crookwellense*, *Fusarium culmorum*, *Fusarium graminearum*, *Fusarium graminum*, *Fusarium heterosporum*, *Fusarium negundi*, *Fusarium oxysporum*, *Fusarium reticulatum*, *Fusarium roseum*, *Fusarium sambucinum*, *Fusarium sarcochroum*, *Fusarium sporotrichioides*, *Fusarium sulphureum*, *Fusarium torulosum*, *Fusarium trichothecioides*, *Fusarium venenatum*, *Humicola insolens*, *Humicola lanuginosa*, *Mucor miehei*, *Myceliophthora thermophila*, *Neurospora crassa*, *Penicillium purpurogenum*, *Trichoderma harzianum*, *Trichoderma koningii*, *Trichoderma longibrachiatum*, *Trichoderma reesei*, or *Trichoderma viride*.

The method of the present invention relies on DNA sequence information to isolate the gene of interest as exemplified elsewhere herein.

Accordingly a preferred embodiment of the invention relates to the method of the first aspect, wherein the sequencing step is done using at least one primer directed to the DNA fragment of the first aspect, or using at least one primer directed to a vector in which the DNA library or cDNA library of the first aspect is cloned.

Further a preferred embodiment of the invention relates to the method of the first aspect, whereisolating the complete gene of interest is done utilizing the DNA sequence information obtained in the sequencing step of the first aspect.

The gene of interest to be isolated by the method of the present invention may encode any polypeptide such as a polypeptide with pharceutical properties, a peptide hormone, an antibody or an antibody fragment, a receptor, or an enzyme.

Consequently a preferred embodiment of the invention relates to the method of the first aspect, wherein the complete gene of interest encodes an enzyme that is secreted from the host cell.

As mentioned previously the method of the invention can be used to isolate a gene of interest to be expressed in an industrial scale later, however this would likely require the con-

struction of an expression system such as described in the art and referenced elsewhere herein.

A preferred embodiment of the invention relates to the method of the first aspect, wherein an additional step of constructing an expression system is performed, said expression system comprising the complete gene of interest isolated in the first aspect.

A gene of interest isolated from a gene library, wherein said gene encodes a protein carrying a secretion signal sequence, and wherein said gene is isolated by the method of the present invention according to the third aspect.

An enzyme encoded by a gene of interest as defined in the third aspect.

An expression system comprising a gene of interest as defined in the third aspect.

A host cell comprising an expression system as defined in the previous aspect.

A host cell comprising at least two chromosomally integrated copies of a gene of interest as defined in the third aspect.

A process for producing a polypeptide comprising cultivating a host cell as defined in the previous aspects under conditions suitable for expressing a gene of interest as defined in the third aspect, wherein said host cell secretes a polypeptide encoded by said gene into the growht medium.

A preferred embodiment of the invention relates to the process of the final aspect, wherein the polypeptide is an enzyme.

Finally a preferred embodiment of the invention relates to the process of the final aspect, where an additional step of purifying the polypeptide is performed.

Examples

Example 1

Construction of a transposon containing the β-lactamase reporter gene *sig*A. This example utilizes a β-lactamase from which the secretion signal has been removed. The β-lactamase conveys ampicillin resistance on *E.coli* only when the protein is secreted to the periplasm, cytoplasmic expression of β-lactamase does not confer ampicillin resistance. Without a signal sequence the β-lactamase enzyme will not be transported to the periplasm and therefore that clone will not grow on media containing ampicillin. A β-lactamase gene is transferred to the target clone using in *vitro* transposition of the transposon described below.

The construction of a transposon containing a signal-less β-lactamase gene was carried out using standard molecular biology techniques. The signal-less β-lactamase gene was initially PCR amplified from commercially available sources (such as from the vector

pUC19) using a proofreading polymerase (Pfu Turbo for example). The resulting PCR fragment contained the restriction sites *Not*I and *Eco*RI in order to aid cloning.

The mini-transposon MuA encoding chloramphenicol resistance was PCR amplified from a commercially available kit (Finnzymes) using a proof reading polymerase (Pfu Turbo) and the primer MuA-F (SEQ ID No.1): 5'-GAAGATCTGAAGCGGCGCACGA. The resulting transposon containing PCR fragment was purified and ligated into the vector pK184 containing a kanamycin resistance gene.

The ligation mixture was electroporated into *E. coli* DH10B and clones containing pK184 with the transposon fragment inserted were selected on LB medium containing chloramphenicol and kanamycin. Many colonies were recovered and plasmid DNA was isolated from 10 of them. Sequencing revealed the correct insertion of the signal-less β-lactamase gene into the transposon MuA contained on the plasmid pK184 (Jobling M.G., Holmes R.K. 1990. Construction of vectors with the p15a replicon, kanamycin resistance, inducible lacZalpha and pUC18 or pUC19 multiple cloning sites. Nucleic Acids Res. 18:5315-5316).

The signal-less β-lactamase gene is contained within the transposon in such a way that there is a continuous open reading frame between the transposon border region (approximately 50 bp in the case of MuA) and the β-lactamase coding region. In this way the modified transposon, when it transposes into a gene encoding a protein that is secreted, can cause an in-frame fusion with the target gene. This results in a fusion gene product that is secreted to the periplasm of *E. coli* and conveys resistance to the ampicillin. Not all transposition events into secreted genes will result in a successful in-frame fusion but when using a positive selection we can screen high numbers and thereby select for even very infrequent events.

Example 2

Use of the transposon sigA containing a signal-less β-lactamase as a reporter gene in the signal trapping of the extracellular xyloglucanase XYG1006.

First the sigA minitransposon is transposed into a cloned subgenomic fragment that contains a known gene encoding an assayable secreted gene-product. In this example we use a xyloglucanase from *Paenibacillus polymyxa*. The xyloglucanase is a large open reading frame (3036 bp) on a subgenomic clone fragment of 4.6 kb in size.

Step 1: Linear mini transposons were prepared by PCR of psigA with Pfu turbo polymerase (Stratagene Inc., USA) using the primer muA-f (SEQ ID 1) amplifying the entire mini transposon. The mini transposons were purified using a GFX column (Pharmacia), diluted to 23ng/ul and used in the standard Finnzyme GPS transposition protocol.

Step 2: The signal trapping mini transposon sigA, the plasmid pXYG1006, 5X buffer and the transposome were mixed in an Eppendorf® tube in the appropriate concentrations and the *in vitro* transposition reaction was performed according to the original Finnzymes protocol. A control experiment using the same plasmid with the original CAM minitransposon was performed in parallel. The transposition reactions were transformed into *E.coli* XL1-blue electrocompetent cells (Stratagene, USA) by electroporation in a Biorad Gene Pulse device (50uF, 25mAmp, 1.8 kV). The cells were diluted in 1ml SOC media and preincubated in a 37°C shaker for one hour. Appropriate dilutions were plated on the LB solid medias listed below to determine the transformation, transposition and signal trapping efficiency.

Solid LB media:

LB-kan (50mg/ml kanamycin).

LB-CAM (10mg/ml chloramphenicol).

LB-CAM-AMP (10mg/ml chloramphenicol, 100mg/ml ampicillin).

LB-CAM, amp, AZCL-xyloglucan (10mg/ml chloramphenicol, 50mg/ml ampicillin, 0.07% w/v AZCL-xyloglucan).

Colonies growing on LB-CAM-AMP were replica plated on LB-CAM-AMP AZCL-xyloglucan to obtain the frequency of disruption of the xyloglucanase domain which is in the first 900 bp of the ORF.

Table 1. Typical results of transposition into pXYG1006

Selection media	Transformants per µg plasmid DNA	
	PSigA	CAM transposome
LB-kanamycin	$3{,}3E10^8$	10^9
LB-CAM	$7{,}5E10^8$	10^6
LB-CAM-AMP	10^4	0
LB-CAM-AMP AZCL xyloglucan	10^3	0

The *E.coli* clones selected on ampicillin and chloramphenicol were those where the β-lactamase reporter gene made a translational fusion with the XYG1006 xyloglucanase gene so that the XYG1006 signal peptide caused the transport of β-lactamase to the periplasm of *E.coli.* Sequencing confirms that all positive clones contain the transposon downstream of the signal sequence. Plasmids from ten random ampicillin resistant colonies are prepared using the Qiaspin procedure (Qiagen) and DNA sequences are determined from the clones using the two primers specific for the transposon:

SigA-r (SEQ ID 2): GCACCCAACTGATCTTCAGCA

SeqB (SEQ ID 3): TTATTCGGTCGAAAAGGATCC

Analysis indicates that SigA lands in the XYG1006 coding region in frame with the xyloglucanase open reading frame.

Example 3

Identification of genes coding for a protein containing a signal sequence in a genomic library using the new transposon sigA. A subgenomic plasmid DNA library is tagged with the signal trapping mini transposon sigA according to the methods described in Example 2. In this example we use a *Paenibacillus pabuli* genomic library prepared by standard methods. The transformation should be plated out on media 1, 2, and 3 (table 2).

Table 2. Typical results of transposition into a *Paenibacillus pabuli* genomic library

Selection media	Transformants per µg plasmid DNA
Medium 1; LB-kan	10^9
Medium 2; LB-CAM	10^6
Medium 3; LB-CAM,amp	100

Plasmid DNA is isolated from positive clones that grow with chloramphenicol and ampicillin (selection medium 3) and can be sequenced from primers that target sequences located in the transposon. In this way the DNA sequence of the signal trapped gene can be obtained. In many cases, single reads with the two transposon primers will yield most of the genetic sequence of the coding region, alternatively custom primers can be synthesized from the sequence obtained in the first run to complete the gene sequence. Another method is to generate 3-100 times more transformants than needed for full coverage of the library. This permits the transposon to land in the same gene but in a different position of the gene within each clone in several independent transposition events. A computer contig assembly program can be adapted to assemble transposants that represent overlapping regions of the same gene. In this way complete or nearly complete coverage of many secreted genes are obtained.

Example 4

Using the information from a signal trapping project. The acquisition of sequence information for all or many of the genes encoding secreted proteins from a gene library is the first step. Most of the trapped genes represent secreted enzymes of known or unknown function. The genes can accordingly be separated into two categories and treated accordingly.

One category of ORFs have significant similarity at the amino acid level to known enzymes. These ORFs can be subcloned into optimal expression vectors, and the constructs

can be used to express significant levels of the enzyme, which can then be tested in various applications.

Another category of ORFs do not have significant homology to any known enzymes but are equally interesting. These can be subcloned into expression vectors and expressed in the same way as the known ORFs. Since however, the enzymatic activity (if any) of these ORFs is unknown, no specific assay exists to monitor their activity, and random application testing is appropriate.

Example 5

Eukaryotic Signal trapping with transposons. Many Eukaryotes also secrete enzymes, fungi for example secrete many classes of enzymes including proteases, cellulases and lipases. Because of the relative size and complexity of eukaryotic genomes, genes encoding enzymes are typically expression cloned from cDNA libraries or are identified in EST (expressed sequence tags) sequencing programs. cDNA libraries are made from mRNA isolated from induced biomass from the eukaryote. Methods are known in the art for representing a broad diversity of secreted enzymes in the cDNA library, these methods include: Pooling of biomass material from separate and different induction conditions followed by normalization of the mRNA or cDNA prior to or after cloning.

The basic theory behind signal trapping in prokaryotes and eukaryotes is essentially the same. The main differences are as follows: cDNA libraries depend on the promoter supplied by the vector into which it is cloned. The cDNA library is a subset of the genome that is expressed which means that the hit rate for the transposon into coding regions is higher than signal trapping from prokaryotic genomic libraries.

The signal trapping marker must be specific for the organism one screens in. Typical screening organisms for fungal genes for example are: *Saccharomyces cerevisiae, Aspergillus niger,* or *Schizosaccharomyces pombe.* In this example we use an invertase signal trapping system described in: Jacobs, K.A., 1997, Gene 198:289-296.

The modified invertase gene is cloned by PCR to include *Not*I and *Eco*RI sites for cloning in frame into the pSigA minitransposon. The bata lactamase is removed by restriction digest and gel purification. A ligation reaction allows the cloning of the invertase gene into the pSigA minitranposon so that the invertase is fused in frame with the left transposon border reading frame exactly as described in the prokaryotic version of pSigA. The completed clone: pSigB is ready for testing in yeast.

The initial test is made on a plasmid containing a cDNA coding for a secreted enzyme that has been expression cloned. The cDNA is the rhgA gene coding for a rhamnogalaturnoase of *Aspergillus aculeatus* (Kofod et al; 1994. J Biol Chem 46:29182-29819). In vitro transposition reactions are performed with 23ng of SigB minitransposon

exactly as described in the bacterial method above. The treated rhgA plasmid is then transformed into yeast cells W3124 in which the native invertase gene is removed. Colonies are plated at high density (1000 colonies per plate) and replica plated on SC media (Sherman, F. 1991. Methods Enzymol., 194:3-21) with sucrose or raffinose.

Table 3. Typical results of transposition into pRhgA

Selection media	**Transformants per µg plasmid DNA**	
	pSigB	
SC media with glucose	1E5	
Replicac plated on SCmedia with sucrose	2E3	

DNA from the yeast colonies capable of growing on sucrose is rescued into *E.coli* by the method of Strathern and Higgens (1991, Methods Enzymol. 194:319-329). Plasmid DNA is isolated with the Qiaspin protocol (Qiagen) and plasmids are sequenced with YES2.0 vector primers and transposon primers to determine the sequence of the insert. In most cases sequence determination with the primers mentioned is sufficient for complete sequence coverage of the cDNA thus allowing analysis of the full length gene and construction of an active expression clone.

Claims

1. A method for identifying and isolating a gene of interest from a gene library, wherein said gene encodes a polypeptide carrying a secretion signal sequence, the method comprising the steps of:

a) providing a genomic DNA library or a cDNA library;

b) inserting randomly into said library a DNA fragment comprising a promoterless and secretion signal-less gene encoding a secretion reporter;

c) introducing the library carrying random inserts of said DNA fragment into a population of host cells;

d) screening for a host cell that expresses and secretes the secretion reporter;

e) identifying the gene of interest into which the secretion reporter was inserted by sequencing the DNA flanking the DNA fragment of step b; and

f) isolating the complete gene of interest from the library of step a).

2. The method of claim 1, wherein step b) is performed *in vitro*.

3. The method of claims 1 or 2, wherein the cDNA or the cDNA library is normalized.

4. The method of any of claims 1 - 3, wherein the genomic DNA library or cDNA library is derived from a microorganism.

5. The method of claim 4, wherein the microorganism is a fungus, a filamentous fungus or a yeast.

6. The method of claim 4, wherein the microorganism is a bacterium.

7. The method of claim 4, wherein the microorganism is an archaeon.

8. The method of any of claims 1 – 7, wherein the DNA fragment of claim 1 is a transposon, preferably a MuA transposon.

9. The method of any of claims 1 – 8, wherein the secretion reporter is a protein which, when secreted from the host cells, allows said cells to grow in the presence of a substance which otherwise inhibits growth of said cells.

10. The method of claim 9, wherein the secretion reporter is a β-lactamase or an invertase.

11. The method of any of claims 1 – 10, wherein the host cells are bacterial.

12. The method of claim 11, wherein the bacterial cells are *Bacillus* cells, preferably *Bacillus alkalophilus*, *Bacillus amyloliquefaciens*, *Bacillus brevis*, *Bacillus circulans*, *Bacillus clausii*, *Bacillus coagulans*, *Bacillus lautus*, *Bacillus lentus*, *Bacillus licheniformis*, *Bacillus megaterium*, *Bacillus stearothermophilus*, *Bacillus subtilis*, or *Bacillus thuringiensis*.

13. The method of any of claims 1 – 10, wherein the host cells are fungal.

14. The method of claim 13, wherein the fungal cells are *Candida*, *Kluyveromyces*, *Pichia*, *Saccharomyces*, *Schizosaccharomyces*, *Yarrowia*, *Acremonium*, *Aspergillus*, *Aureobasidium*, *Cryptococcus*, *Filibasidium*, *Fusarium*, *Humicola*, *Magnaporthe*, *Mucor*, *Myceliophthora*, *Neocallimastix*, *Neurospora*, *Paecilomyces*, *Penicillium*, *Piromyces*, *Schizophyllum*, *Talaromyces*, *Thermoascus*, *Thielavia*, *Tolypocladium*, or *Trichoderma*.

15. The method of claim 13, wherein the fungal cells are *Saccharomyces cerevisiae*, *Aspergillus aculeatus*, *Aspergillus awamori*, *Aspergillus nidulans*, *Aspergillus niger*, or *Aspergillus oryzae*.

16. The method of any of claims 1 – 15, wherein the sequencing step in claim 1 is done using at least one primer directed to the DNA fragment of claim 1, or using at least one primer directed to a vector in which the DNA library or cDNA library of claim 1 is cloned.

17. The method of any of claims 1 – 16, where isolating the complete gene of interest is done utilizing the DNA sequence information obtained in the sequencing step of claim 1.

18. The method of any of claims 1 – 17, wherein the complete gene of interest encodes an enzyme that is secreted from the host cell.

19. The method of any of claims 1 – 18, wherein an additional step of constructing an expression system is performed, said expression system comprising the complete gene of interest isolated in claim 1.

20. A gene of interest isolated from a gene library, wherein said gene encodes a polypeptide carrying a secretion signal sequence, and wherein said gene is isolated by the method of the present invention.

21. An enzyme encoded by a gene of interest as defined in claim 20.

22. An expression system comprising a gene of interest as defined in claim 20.

23. A host cell comprising an expression system as defined in claim 22.

24. A host cell comprising at least two chromosomally integrated copies of a gene of interest as defined in claim 20.

25. A process for producing a polypeptide comprising cultivating a host cell as defined in claim 23 or 24 under conditions suitable for expressing a gene of interest as defined in claim 20, wherein said host cell secretes a polypeptide encoded by said gene into the growht medium.

26. The process of claim 25, wherein the polypeptide is an enzyme.

27. The process of claim 25 or 26, where an additional step of purifying the polypeptide is performed.

Abstract

The present invention allows to screen previously established genebanks or libraries by proxy for genes encoding secreted polypeptides or enzymes even of unknown activity and thus without known screening assays, polypeptides that may turn out to be of industrial interest and which would likely not have been isolated using conventional screening assays. A method for isolating genes encoding secreted polypeptides from existing gene libraries is described in which the endogenous secretion signal sequences are detected using an *in vitro* transposition reaction where the transposon contains a secretion reporter.

Figure 1

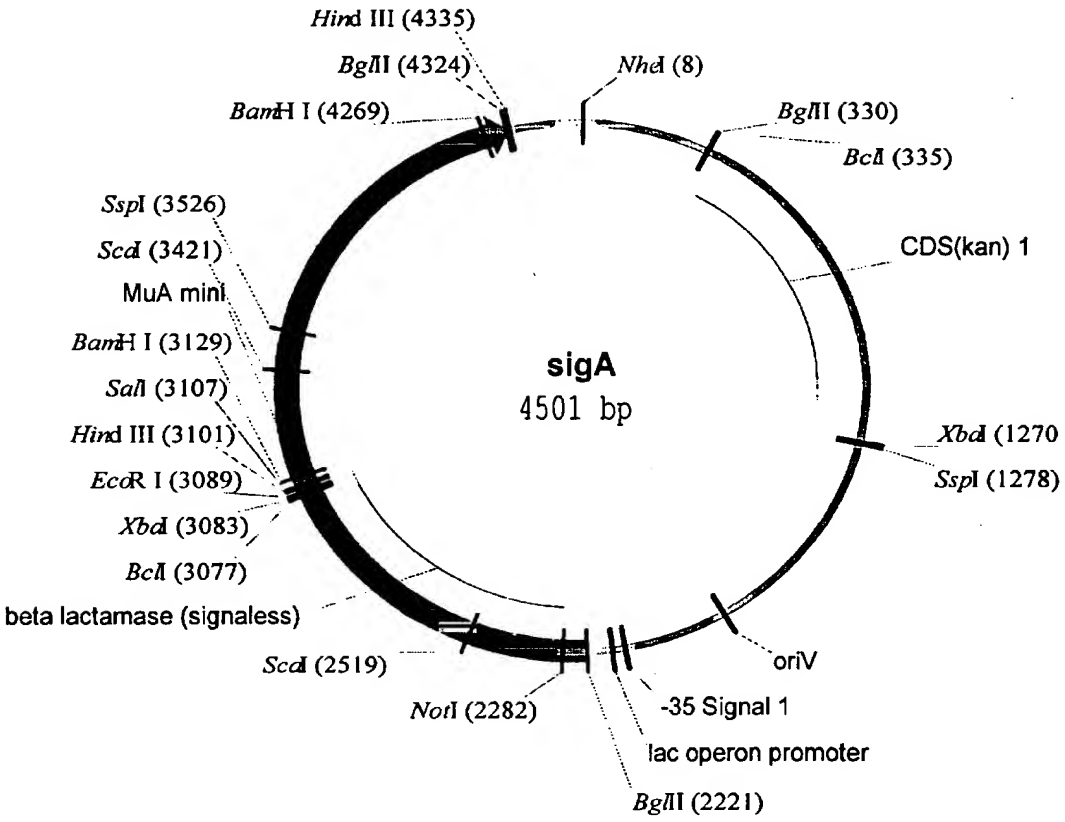
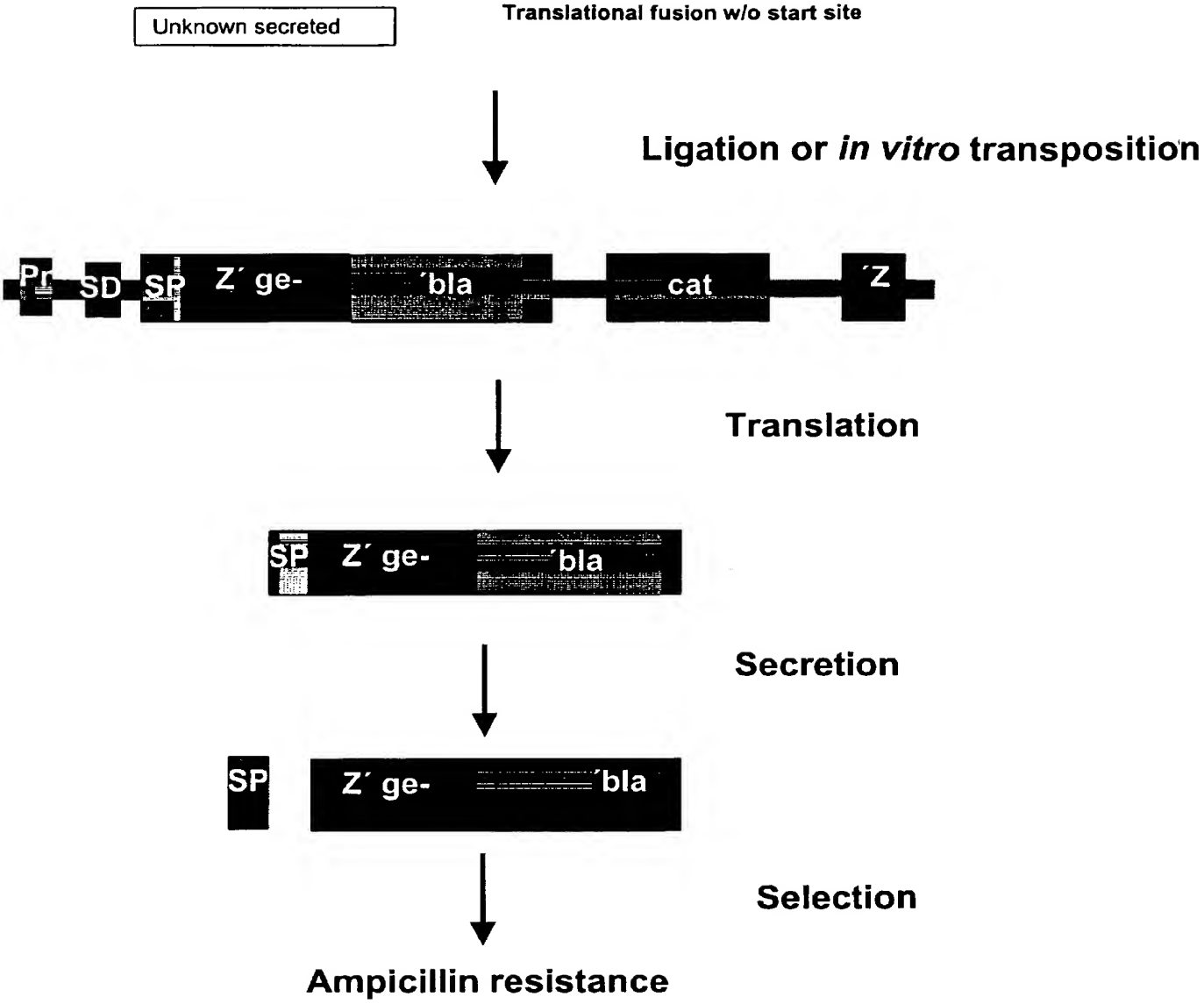

Hind III (4335)
BglII (4324)
BamH I (4269)
NheI (8)
BglII (330)
BclI (335)
CDS(kan) 1
XbaI (1270
SspI (1278)
oriV
-35 Signal 1
lac operon promoter
BglII (2221)
NotI (2282)
ScaI (2519)
beta lactamase (signaless)
BclI (3077)
XbaI (3083)
EcoR I (3089)
Hind III (3101)
SalI (3107)
BamH I (3129)
MuA mini
ScaI (3421)
SspI (3526)
sigA
4501 bp

Figure 2


Unknown secreted
Translational fusion w/o start site
Ligation or *in vitro* transposition
Pr
SD
SP
Z´ ge-
´bla
cat
´Z
Translation
SP
Z´ ge-
´bla
Secretion
SP
Z´ ge-
´bla
Selection
Ampicillin resistance

SEQUENCE LISTING

<110> Novo Nordisk A/S

<120> Signal Sequence Trapping

<130> 10018.000-DK

<140>
<141>

<160> 3

<170> PatentIn Ver. 2.1

<210> 1
<211> 22
<212> DNA
<213> Artificial Sequence

<220>
<223> Description of Artificial Sequence: Primer MuA-F

<400> 1
gaagatctga agcggcgcac ga 22

<210> 2
<211> 21
<212> DNA
<213> Artificial Sequence

<220>
<223> Description of Artificial Sequence: Primer SigA-r

<400> 2
gcacccaact gatcttcagc a 21

<210> 3
<211> 21
<212> DNA
<213> Artificial Sequence

<220>
<223> Description of Artificial Sequence: Primer SeqB

<400> 3

ttattcggtc gaaaaggatc c 21